Regression Final Project World-Happiness Report

Dhruv Shah ds2268@scarletmail.rutgers.edu Kaggle Name: Dhruv

Abstract:This project aims to predict the happiness levels of various countries based on multiple socioeconomic factors over time. The dataset includes variables such as the country, year, happiness, log of GDP per capita, social support, life expectancy, freedom of choices, generosity, corruption, positive affect, and negative affect. The primary objective of this study is to analyze how these factors contribute to the variation in happiness scores across different nations and over multiple years.

Methodology

- The methodology involves several key steps, starting with data cleaning and preprocessing. Missing values were handled, and outliers were identified and removed to ensure the accuracy of the analysis.
- Feature selection techniques were applied to identify the most significant predictors of happiness.
- The data was then split into training and test sets to build a predictive model using regression techniques. Various statistical analyses were conducted to evaluate the relationships between happiness and the selected variables.

Key Findings:

1. The analysis revealed that key factors such as social support, life expectancy, and freedom of choice had the most significant positive impact on happiness scores.
2. Corruption was found to have a negative correlation with happiness, while positive affect and social support also contributed strongly to higher happiness levels across countries.
3. GDP has a strong positive correlation with happiness, indicating that wealthier countries report higher happiness levels.
4. Social_support variable was crucial in predicting happiness, emphasizing the importance of community and social networks.
5. Higher life expectancy was associated with increased happiness, underscoring the role of health in overall well-being.
6. Freedom & Generosity both factors positively influenced happiness, highlighting the importance of personal freedoms and altruism.

Visualization:

- Various visualizations were created using Matplotlib to illustrate the relationships between happiness and the predictor variables, including feature importance plots to highlight the most significant predictors.

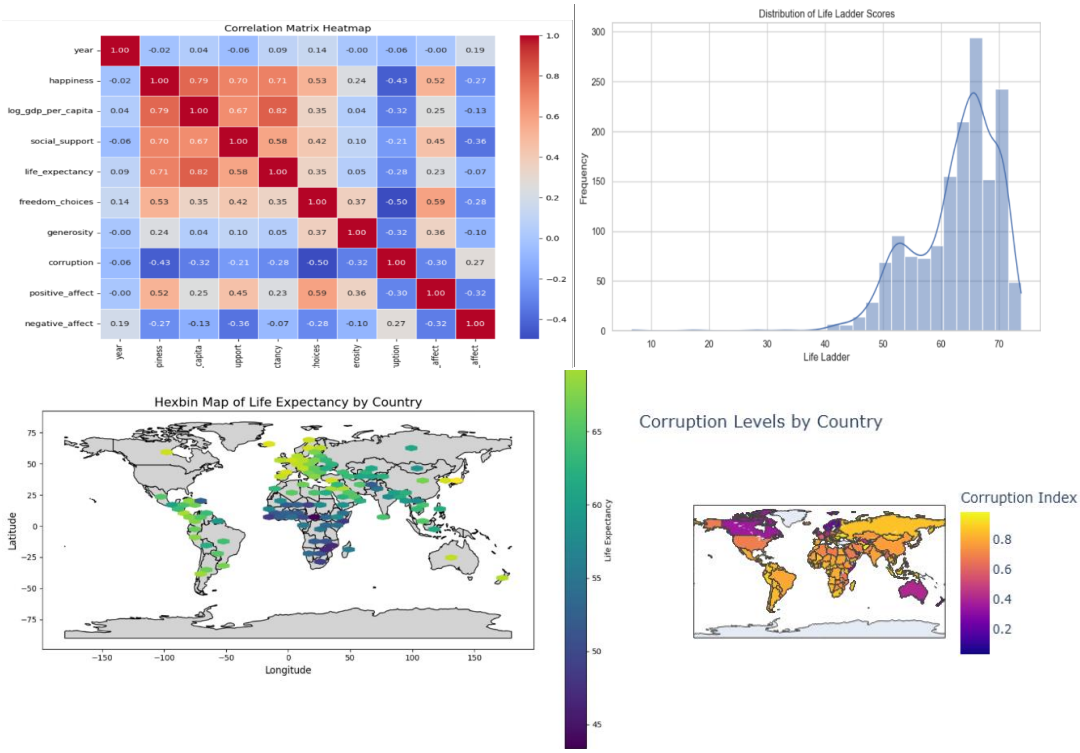
Regression Analysis:

- The Random Forest model was trained on the pre-processed dataset, utilizing the selected socio-economic indicators as predictors for happiness.

- Model performance was evaluated using metrics such as R-squared and mean-squared error, demonstrating the effectiveness of the Random Forest approach.

Discussions:

1. The regression analysis assumes a linear relationship between the predictors and the dependent variable (happiness). If the true relationship is non-linear, the linear regression model may not capture the complexity of the data, leading to biased predictions.
2. If the predictors are highly correlated with each other they achieve multicollinearity. It can distort the result of the regression analysis. This can make it difficult to determine the individual effect of each predictor of the happiness score, potentially leading to misleading conclusions.
3. The interpretation of the results should be approached with caution. Correlation does not imply caution, and while the analysis may identify relationship variables, it does not establish that changes in the predictors directly cause changes in happiness scores.



Linear Regression :The formula for Linear Regression involves finding the line that minimizes the error between predicted and actual values of the dependent variable.

The general equation for **Linear Regression** prediction is:

$$y_{\text{pred}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Y_{pred} is the predicted value of the target (happiness), x1,x2,...,xp are the predictor variables (e.g., corruption, freedom_choices, positive_affect, etc.),

```
Validation Mean Squared Error: 0.2998380718389478
```

Random Forest

- Random forest is a **non-parametric** method that builds an ensemble of decision trees.
- Each tree in the forest is trained on a random subset of the data, and predictions are made by averaging the predictions from each tree.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2$$

Y_{true},i is the true value of the target variable for the i-th data point, y_{pred},i is the predicted value from the random forest model for the i-th data point.

```
Validation Mean Squared Error: 0.1870370590338059
```

Model Selection:

- A comparison was made between Linear Regression and Random Forest Regression models.
- The Random Forest model was chosen due to its superior prediction performance, as indicated by higher accuracy metrics compared to the Linear Regression model.

References

https://vdsbook.com/ (see Exercise 22, Chapter 9, Part III)
https://github.com/Yu-Group/vds-book-supplementary/tree/main/R/exercises/happiness
Report: https://worldhappiness.report/ed/2018/